Inferring gender of a Twitter user using celebrities it follows

Puneet Singh Ludu
State University of New York
Buffalo, New York
pludu@buffalo.edu

arXiv:1405.6667v1 [cs.IR] 26 May 2014

ABSTRACT

This paper addresses the task of user gender classification in social media, with an application to Twitter. The approach automatically predicts gender by leveraging observable information such as the tweet behavior, linguistic content of the user's Twitter feed and the celebrities followed by the user.

This paper first evaluates linguistic content based features using LIWC dictionary and popular neighborhood features using Wikipedia and Freebase. Then augments both features which yielded a significant increase in the accuracy for gender prediction. Results show that rich linguistic features combined with popular neighborhood prove valuables and promising for additional user classification needs.

Keywords

Social Information retrieval

1. INTRODUCTION

Twitter a micro-blogging site, have become an integral part of the life of millions of users. People use it to communicate with friends,family or acquaintances. With the increase in the number of users, demand for analytics on this data is also growing. Twitter does not store gender, age, ethnicity or interests of a user. These attributes of a user could be useful both for user experience as well as for consumption by brands in their social analytics.

In this work we address the task of user gender classification, by leveraging observable information such as the tweet behavior, linguistic content of the user's Twitter feed and the celebrities followed by the user.

Main contributions of this work are following:

- It uses some new linguistic content based features, which were not used by state-of-the-art, particularly on Twitter data.
- It describes completely new set of popular(celebrity) neighborhood based features such as *age, gender and occupational area of the celebrity* which show promising results.
- It reports that using combination of both features can perform better than various state-of-the-art techniques, and especially extraction of better popular network based features demands further investigation.

The paper is organized as follows. Section 2 introduces relevant and related work on user profiling for social media, Twitter user attribute detection and topic models for Twitter. Section 3 describes the datasets we have used or created. Section 4 we explains methodology we followed to clean and process the datasets, while Section 5 explains in detail use of various features and the intuition behind their usage. Section 6 describes results produced in different configurations. Finally, Section 7 draws final conclusions and outline future work.

2. BACKGROUND AND RELATED WORK

Inferring attributes of social media users is a growing area of interest, there have been many recent attempts to predict various hidden attributes of a user. Several of the recent works were focused on predicting ethnicity (Rao et al., 2011; Pennacchiotti and Popescu, 2011), age (Schler et al., 2006; Rosenthal and McKeown, 2011; Nguyen et al., 2011; Al Zamal et al., 2012), gender (Rao et al., 2010; Burger et al., 2011; Liu and Ruths, 2013; Al Zamal et al., 2012), Interests (Lim and Datta, 2013), personality (Argamon et al., 2005; Schwartz et al., 2013) etc.

Most of the datasets prepared in these approaches were hand annotated, cherry picked or attributes are identified by the user itself. Also many of the datasets were limited to very broad attribute categories, such as age below 23 as attribute young and above 25 as old. Rao et al. (2011) used linguistic features to predict ethnicity and gender of facebook users. Using a very limited training data, they tried to evaluate fine grained ethnicity classes of Nigeria. Social or Network based features were first used by Pennacchiotti and Popescu (2011); they tried to predict if a Twitter user is African-American or not.

As preparing a good dataset is a big challenge, Schwartz et al. (2013) tried to explore this area and collected Facebook profiles labeled with personality type, gender, and age by administering a survey of users embedded in a personality test application.

Al Zamal et al. (2012) explored into social features by exploring homophily based features; such as using linguis-

tic features of n-most-popular friends of a user on Twitter. Their final classifier obtains a **80.02%** accuracy on binary gender classification (compared with the **85.67%** accuracy this paper report below for same dataset).
Lim and Datta (2013) presented a novel approach for classifying user interests for e.g. Food, Politics, Charity etc. using celebrity(popular users) followed by a twitter users.

3. DATASETS

In order to collect the datasets, it was necessary to decide on the two contrasting labels (for e.g "male" and "female") that would be used and identify users that could be reliably assigned one label or the other. Even though we have used dataset which was already prepared by Zamal, still we had to manually inspect accounts collected to confirm that the labeled assignments were correct.

3.1 Gender Dataset

This paper have used the same dataset prepared by Al Zamal et al. (2012) for gender prediction. Zamal's paper used a technique proposed in Mislove et al. (2011), which used users who had their full first and last names on Twitter; also their first name was one of the top 100 most common names on record with US social security department for baby boys/girls born in the year 2011. 131 male and 134 female labeled users were collected out of 192 each, because some users closed their Twitter account or locked their timeline for public access.

3.2 Popular neighborhood Dataset

We have also created a dataset of the popular neighborhood or so called celebrity users followed by these 265 users. Here, we define celebrities(popular users) as users with more than 10,000 followers or who have been verified by Twitter itself, as was defined by Lim and Datta (2013). Thus, out of 108,619 distinct users followed by 265 users, we were left with 5,546 popular users. Once we had the list of these celebrity users we extracted various features which we would discuss in section 5.

4. METHODOLOGY

The initial results of applying user tweets based classification did not seem very promising, a closer examination revealed that the problem was in the quality of the Tweet text. Since the text in this domain comes from Short messages, and the messages often contain various abbreviations, and are rife with spelling errors. Thus, to deal with this, we implemented a process to auto-correct words that are greater than 3 letters in length as suggested by Prem Melville et al. (2013). This process leveraged a combination of Philips' metaphone algorithm and string-edit(TextBrew) distance for spell-correction; in addition to a general-purpose English dictionary, we used dictionary for Twitter short hands, for e.g. *"idk"* would be mapped to *"I don't know"*. As seen on Twitter many people use a lot of camel-casing especially while using hashtags, we tried to clean such camel-cases, for e.g. *"#BestDayEver"* would be cleaned as *"Best Day Ever"*. After cleaning tweets we saw 12% gain in accuracy only based on user tweets based classification.

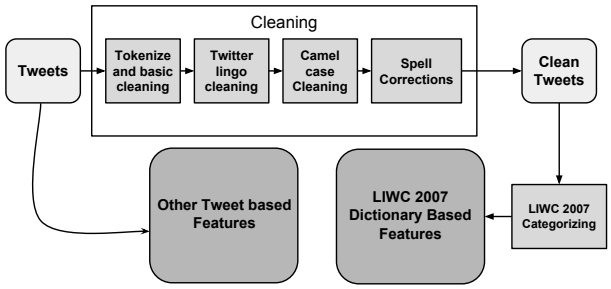

Figure 1: High level view of tweet based features extraction

5. FEATURES

Previous works in this area have already explored some general lexicon features based on tweet text for e.g. K-top words used by each class. In this paper we have not used these features, instead we used LIWC 2007 dictionary for extracting lexicon features. With LIWC based lexicon features we got almost 3% gain in accuracy compared to state-of-the-art lexicon features as shown in Section 6.
We will now discuss all the features used in this paper for predicting gender of Twitter user.

5.1 Tweet behavior based features

Tweeting behavior can be a good parameter to distinguish between male and female, we have used almost all the features used by state-of-the-art.

5.1.1 Tweet Frequency

Let's say number of tweets extracted for a user be N (in this paper we used N=1000) and chronological difference in days between first and last tweet be C, then "tweet frequency" is defined as

$$TF = \frac{N}{C} \tag{1}$$

According to a study done by *www.beevolve.com* on 36 million twitter users in 2012, it was found that on average female users tend to tweet more often than male counterparts.

5.1.2 Hashtag Frequency

Let T_h be the total number of hashtags used by a user, then "hashtag frequency" is defined as

$$HF = \frac{T_h}{C} \tag{2}$$

5.1.3 Average Tweet Length

"Average Tweet Length" TL represents the average length of tweets. Here, TL is defined as

$$TL = \frac{\sum_{i=0}^{RecentN} Length(tweet_i)}{N} \tag{3}$$

5.1.4 Retweet Frequency

Let T_r be the total number of tweets retweeted by a user, then "retweet frequency" is defined as (Rao et al. 2010):

$$RF = \frac{T_r}{C} \tag{4}$$

5.1.5 Followers to following ratio

The ratio between number of followers and number of friends has been used as a measure of a user's tendency towards producing vs. consuming information on Twitter (Rao et al. 2010).

5.1.6 Celebrity following tendency

Let T_p be the total number of celebrities or popular users followed by a user and T_f be the total number , then 'celebrity following tendency" is defined as

$$CT = \frac{T_p}{C} \tag{5}$$

5.2 Linguistic content based features

Linguistic content information encapsulates the main topics of interest to the user as well as the user's lexical usage. In a study done by *www.beevolve.com* (2012), Female users talk more about family and fashion whereas the Twitter male users prefer technology, sports and entrepreneurship. We explored a wide variety of linguistic content features using LIWC[1] 2007 dictionary, as detailed below.

5.2.1 Count of *I, we, you, he-she* and *they* words

According to David Bamman et al. (2014), female Twitter user tend to use more personal diary writing style, where they might use more *I* and *he-she* references.

5.2.2 Parts of Speech

Pronouns are generally associated with female authors (David Bamman et al., 2014), **Conjunction** such as *and* were associated with female authors. Other parts of speech such as articles, determiners or prepositions showed low gender association.

5.2.3 Emotions and Emoticons

Emotional terms such as sad, love, glad, etc. are more associated with female authors, while female author show more associativity with emotionally neutral sentences (David Bamman et al., 2014).
We have also used Emoticons such as :-), <3 etc. and their unicode counterparts. Emoticons tend to have more associativity with female authors.

5.2.4 Categorical words

Words related to Health, Money, Achievement, Society, Family etc. can be a good indicator for predicting gender of a twitter users. As shown in many studies words related to "Money and" "Finance" are often used by male authors, while words related to "Family", "Kinship" and "Society" are more often used by female authors on Twitter. Similarly "Swear" words are more often used by male authors rather than female authors (David Bamman et al., 2014; *www.beevolve.com*, 2012). In this papers we have used 50 such categories, each as one feature.

[1]Linguistic Inquiry and Word Count (LIWC) is a text analysis software program designed by James W. Pennebaker, Roger J. Booth, and Martha E. Francis. LIWC calculates the degree to which people use different categories of words across a wide array of texts, including emails, speeches, poems, or transcribed daily speech.

Categories	***Associativity***
Pronouns	Female
Emotion terms	Female
Kinship terms	Female
CMC words (lol, omg)	Female
Numbers	Male
Technology words	Male
Swear words	Male
Assent	Female
Emoticons	Female
Hesitation	Female

Table 1: Some of the categories and their known associativity with gender of user (David Bamman et al., 2014)

In total we are using 64 different Linguistic content based features. These features proved to give better results than other simple Linguistic content based features such as *K-top* words.

5.3 Popular neighborhood based features

As we have already defined "popular neighborhood" or "celebrity users" as those users who are followed by any of the 265 users in our dataset having either (1) more than 10,000 followers or (2) are verified users[2]. Popular users (celebrities) use their actual name on Twitter and it is feasible to extract their features from websites like Wikipedia and Freebase. This gives us an opportunity to leverage principle of homophilic association (in Twitter, a user can exercise greater selectivity over who she follows than who follows her) with greater accuracy in extracting neighborhood features.
In the following subsections we would discuss our intuition behind each "popular neighborhood based feature" we have chosen and how we extracted those features.

5.3.1 Age of celebrity

We have used 5 categories(each as a feature) for a celebrity users age

- age < 23 years
- 23<age<30
- 30<age<40
- 40<age<50
- 50<age

for e.g. Justin Bieber would be of "age<23" category. The basic intuition behind taking celebrity age as feature for predicting gender is, number of female users under age 30 years on Twitter outnumbered males under age 30, this gap even widens under age 20. Concept of homophilic association suggests that users of same age category tend to follow each other more. Thus, probability of female following celebrities under age 23 years would be more than male. We used

[2]In June 2008, Twitter launched a verification program, allowing celebrities, brands, businesses and public figures to get their accounts verified, there are currently approximately 92,000 Verified users on Twitter

Wikipedia to extract age of the celebrity. Later in our experiment with this feature, we found that age of a celebrity has a weak associativity with the gender of its follower.

5.3.2 Gender of celebrity

It has been discussed by Bill Heil et al. (2009), that an average man is almost twice more likely to follow another man than a woman. We have used "male" and "female" category of celebrity as feature(for e.g. Sachin Tendulkar would be categorized as male), and ignored gender neutral Twitter users such as Brand and Business accounts.

	a female	*a male*
Female follows	44%	56%
Male follows	35%	65%

Table 2: Tendency of male and female following other male and female (Bill Heil et al., 2009)

Since it is almost impossible to find gender of the complete neighborhood of a user, we have used just popular neighborhood. We extracted gender of a celebrity using Wikipedia by counting *he, she, her, his, him etc.* references on that celebrities page. In our experiment we found promising evidence suggesting gender of neighborhood could be helpful in predicting gender of user itself.

5.3.3 Celebrities "famous for"

Work of Lim and Datta (2012) suggests that celebrities represent an interest category, and they leveraged these interests to discover communities who follow specific interests on Twitter. In this paper we have also used similar categories to find interests of a user using the kind of celebrity they follow. We have used followed following categories each as one feature.

"famous for" Categories of celebrities	
• Acting	• Art
• Entertainment	• Entrepreneur
• Writing	• Music
• Politics	• Religious
• Science & Technology	• Security
• Social	• Sports
• Miscellaneous	

According to a study conducted by *www.beevolve.com* on 36 million twitter users, it was found that some of the broad categories were more distinctive then other. And these are of interests could be used predict gender and other attributes of a user.

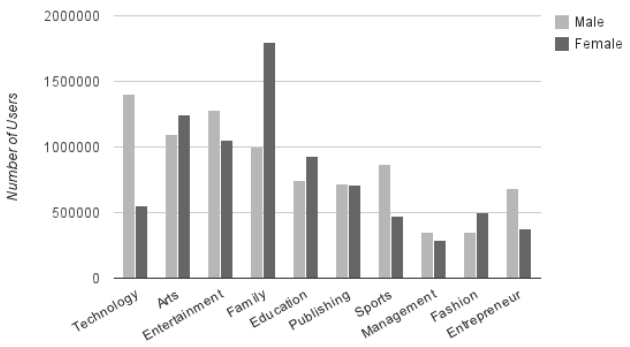

Figure 2: Gender distribution by top 10 categories of interest on Twitter (*www.beevolve.com*, 2012)

6. RESULTS

We would evaluate our results in terms of overall accuracy, A 10-fold cross-validation was done to assess the performance of the SVM-based classifiers at inferring the gender of a Twitter user.
We will discuss the results in following three stages:

6.1 Using tweet behavior and linguistic features

Configuration	***Accuracy***
LIWC only(with last 200 tweets)	77.95%
LIWC only(with last 1000 tweets)	80.43%
LIWC + Tweet Behavior features	**82.26%**

Table 3: The overall accuracy while using all the tweet based features which excludes homophily features

6.2 Using popular neighborhood features

Configuration	***Accuracy***
Age of celebrity	56.05%
Celebrity "famous for"	67.23%
Gender of celebrity	**71.57%**
Combined	66.84%

Table 4: The overall accuracy using popular neighborhood features

6.3 Using all the features combined

	True Male	*True Female*	**Precision**
Pred. Male	115	17	87.12%
Pred. Female	21	112	84.21%
Recall	84.56%	86.82%	

Overall accuracy = **85.67%**

Table 5: The final confusion matrix and overall accuracy using all the features

7. CONCLUSION AND FUTURE

In this paper, we evaluated the extent to which features present in a Twitter user's popular neighbors can improve the inference of attributes possessed by the user herself. Our results support several noteworthy conclusions, which we discuss here.

7.1 Usefulness of linguistic features

Liguistic features such as associativity of a particular category of words towards the gender of a user as suggested by David Bamman et al. (2014), produced satisfactory results.

7.2 Popular neighborhood is useful and feature rich

Pennacchiotti et. al (2011), discussed in their paper that users following more celebrities tend to give better results in their homophily(network) based approach. We verified in this paper that popular neighborhood can be leveraged to extract various rich and accurate features using already available knowledge bases.

We have observed that gender of a celebrity can be a good feature, and proves the study conducted by Bill Heil et al. (2009), which suggests tendency of males to follow male is twice than following female.

Information regarding popular user's domain of work can be useful for determining users interest which in turn can be useful in gender classification.

In this paper we have not used other homophily features such as linguistic features of popular neighborhood, it would interesting to see combined results of homophily and celebrity based features used in this paper. We can also explore the possibility of using these features to infer other user attributes such as "age" and "political orientation"

8. ACKNOWLEDGMENTS

I would like to thank Rahul Tejwani for his help and active participation in various discussions for this work. I would also like to thank Professor Rohini Srihari and all those who helped me throughout this project.

9. REFERENCES

[1] F. Al Zamal, W. Liu, and D. Ruths. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *ICWSM*, 2012.

[2] S. Argamon, S. Dhawle, M. Koppel, and J. Pennebaker. Lexical predictors of personality type. 2005.

[3] D. Bamman, J. Eisenstein, and T. Schnoebelen. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160, 2014.

[4] Beevolve.com. An exhaustive study of twitter users across the world. 2013.

[5] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309. Association for Computational Linguistics, 2011.

[6] B. Heil and M. Piskorski. New twitter research: Men follow men and nobody tweets. *Harvard Business Review*, 1, 2009.

[7] K. H. Lim and A. Datta. Finding twitter communities with common interests using following links of celebrities. In *Proceedings of the 3rd international workshop on Modeling social media*, pages 25–32. ACM, 2012.

[8] W. Liu and D. Ruths. What's in a name? using first names as features for gender inference in twitter. In *Analyzing Microtext: 2013 AAAI Spring Symposium*, 2013.

[9] P. Melville, V. Chenthamarakshan, R. D. Lawrence, J. Powell, M. Mugisha, S. Sapra, R. Anandan, and S. Assefa. Amplifying the voice of youth in africa via text analytics. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1204–1212. ACM, 2013.

[10] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist. Understanding the demographics of twitter users. *ICWSM*, 11:5th, 2011.

[11] D. Nguyen, N. A. Smith, and C. P. Rosé. Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 115–123. Association for Computational Linguistics, 2011.

[12] M. Pennacchiotti and A.-M. Popescu. A machine learning approach to twitter user classification. In *ICWSM*, 2011.

[13] J. W. Pennebaker, M. E. Francis, and R. J. Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001, 2001.

[14] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44. ACM, 2010.

[15] S. Rosenthal and K. McKeown. Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 763–772. Association for Computational Linguistics, 2011.

[16] J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 6, pages 199–205, 2006.